# User Modeling Based on Emergent Domain Semantics

Marián Šimko and Mária Bieliková

Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technology, Slovak University of Technology,
Ilkovičova 3, 842 16 Bratislava, Slovakia
{simko,bielik}@fiit.stuba.sk

**Abstract.** In this paper we present an approach to user modeling based on the domain model that we generate *automatically* by resource (text) content processing and analysis of associated tags from a social annotation service. User's interests are modeled by overlaying the domain model – via keywords extracted from resource's (text) content, and tags assigned by the user or other (similar) users. The user model is derived automatically. We combine content- and tag-based approaches, shifting our approach beyond flat "folksonomical" representation of user interests to involve relationships between both keywords and tags.

**Keywords**: user modeling, emergent domain semantics, automatic domain model composition, folksonomy, text mining

## 1 Introduction and Related Work

Recommendation in social systems, also referred to as collaborative filtering, consists of *(i)* user similarity computation and *(ii)* relevant resource prediction. The purpose of the first step is to find the most similar users with the "active" user, often assuming their similar behavior during the process of information search or navigation (visiting a page, buying a product). Similar behavior is interpreted as similar interest, which is a base for the second step, where resources (pages, products) are predicted based on their relation to the most similar users.

Traditional approaches to the user similarity computation utilize methods of usage mining [9]. Visiting the same web page or similar movie rating indicates similar interest. The other group of approaches is based on social tagging. Tags are promising source of information for recommendation as the number of tagging users all along increases. From the user modeling perspective, tags represent user interest. Strictly speaking, the action of assigning a tag to a resource is what is interpreted as user interest in tagged resource [2, 3]. Furthermore, different tags from different users are analyzed to consider the context of tagging [7, 8]. In order to derive more accurate recommendation, contextualized score for resources is being computed for each user to reveal different purposes of tags.

In our work we primarily focus on recommendation of text-based resources such as web pages or learning objects in collaborative learning environment. In

the thesis proposal we combine social-based collaborative filtering with content-based approach. We utilize tags from tag-based systems, but we shift the whole approach beyond flat folksonomical representation of resources, leveraging light-weight emergent semantics generated automatically based on resource analysis.

## 2 Emergent semantics

When selecting users whose associated resources will be recommended, their models are compared in order to obtain user similarity level. User model composition is one of the most delicate parts of any method for personalized search or recommendation. We build on overlay user model that is based on the domain model [1]. The crucial part of our work and our contribution is automated user model generation based on the resource content and tag analysis. The acquired representation we refer to as resource metadata (see Fig. 1).

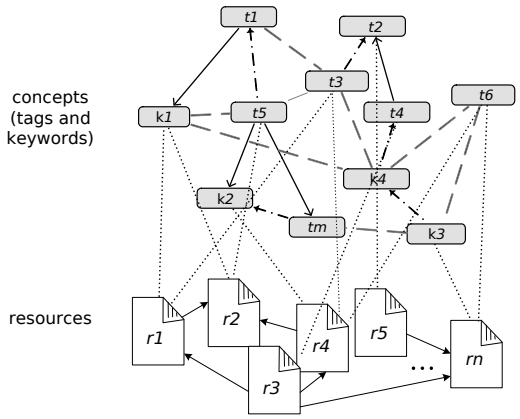


**Fig. 1.** User model representation (upper part) for a set of resources (pages) he visited. Entity k$x$ represents keyword, t$y$ represents tag and r$z$ represents resource. Different relationship markup between entities reflects different semantic power.

Metadata consist of *concepts*[1] and *relationships* between concepts. We differentiate two types of concepts:

- keywords (content),
- tags (folksonomy).

Concepts represented by keywords extracted from resource content have different semantics than tags assigned by a user himself. While the first type of concept is

[1] Using term *concept* can be slightly misleading as someone can think of the concept only for representing conceptual knowledge, abstract or general ideas inferred or derived from specific instances. We use term *concept* because resource metadata serve exactly by the same way as conceptual knowledge.

added to the user model when visiting a page (similarly to [5]), the second type is added when the user tags a page. The latter action is more explicit and reflects into the higher weight of relation between a resource and a tag when computing concept and user similarities. We believe that considering tags together with the resource content is feasible, as research of social tagging showed that tags are to a certain, significant, extent dependent on resource's content (e.g. title) [4].

Relationships between concepts represent concepts' relatedness. We create them automatically by an underlying graph analysis utilizing the notion of node centrality. As they are derived from underlying domain model, they can be viewed as emergent domain semantics representing additional value to pure concept-based user model. The method for relationship generation we already evaluated in e-learning domain when discovering relationships between concepts extracted from learning objects [11]. Similarly to relationships, keywords are extracted automatically. We apply here the methods and techniques for automated term recognition [6].

Based on described representation, user similarity we compute considering following principles: users are more similar if

- the more similar concepts are assigned to resources they visit;
- the more similar tags are assigned (by other users) to resources they visit;
- the more similar tags they assign to same resources;
- the more similar relationships between concepts (keywords and tags) exist.

The user model is generated automatically and it is different for every user (it contains different concepts and different relationships between them, which are derived from users' actions). The user's context is considered as we track the way he accesses the resource: by visiting and/or by tagging.

The next step, resource prediction, is based on the user model similarity computation. For the most similar users (those exceeding a certain similarity threshold), a prediction score is computed. We consider two computational variants: statistical and topological, each representing different view on two user models. Variants can be mutually combined in order to achieve better recommendation.

## 3 Conclusions

In our work we focus on automatic composition of a user model. We proposed the method that builds the user model *combining* the content-based and tag-based approach. After a resource's content and assigned tags are analyzed, concepts represented by keywords and tags are added to user model. The relationships between them representing relatedness of entities are composed considering the user's context. This approach building on the domain model created automatically we consider the main contribution that the thesis aims to achieve. The created user model is used for user similarity computation and resource prediction computation for recommendation.

In the current stage of our research we have analyzed methods for automatic term extraction, we have analyzed methods for relationship discovery from the

text (as a part of *ontology* learning field), and we have analyzed methods for concept relationship induction from folksonomies. We proposed a method for automatic relationship discovery based on underlying graph representation (the graph on Fig. 1 with no relationships between concepts yet) that we evaluated in the e-learning domain [10].

As important research questions we would like to discuss the following: What is the relevance of relationships in the user model from user modeling perspective and to what extent they would be different if created manually? What specific resource content (text) features can be exploited to fine tune the user model and improve accuracy of extraction of keywords more suitable for user model?

*Acknowledgments.* This work was partially supported by the Scientific Grant Agency of SR, grant No. VG1/0508/09, the Cultural and Educational Grant Agency of SR, grant No. 028-025STU-4/2010 and by the Research & Development Operational Programme for the project Support of Center of Excellence for Smart Technologies, Systems and Services II, ITMS 26240120029, co-funded by the ERDF.

## References

1. Brusilovsky, P. Methods and Techniques of Adaptive Hypermedia. In: User Modeling and User-Adapted Interaction, 6 (2-3), 1996, pp. 87-129.
2. Carmagnola F., Cena, F., Cortassa, O., Gena, C., Torre, I. Towards a Tag-Based User Model: How Can User Model Benefit from Tags? In: Conati, C., McCoy, K., Paliouras, G. (eds.): LNAI 4511: User Modeling 2007, Springer-Verlag, pp. 445-449.
3. Durao, F., Dolog, P. A personalized tag-based recommendation in social web systems. In: Tasso, D. A., Farzan, R., Kleanthous S., Bueno Vallejo, D., Vassileva, J. (eds.): Proc. of Int. Workshop on Adaptation and Personalization for Web 2.0 (AP-WEB 2.0 2009) at UMAP2009, CEUR, vol. Vol-485, Trento, Italy, 2009.
4. Heymann, P., Koutrika, G., Garcia-Molina, H. Can Social Bookmarking Improve Web Search? In Proc. of the WSDM 2008, pp. 195-206.
5. Kajaba, M., Návrat, P., Chudá, D. A Simple Personalization Layer Improving Relevancy of Web Search. In Computing and Information Systems Journal. 13 (3), 2009, pp. 29-35.
6. Knoth, P., Schmidt, M., Smrž, P., Zdráhal, Z. Towards a Framework for Comparing Automatic Term Recognition Methods, In: Znalosti 2009, pp. 83-94.
7. Nakamoto, R., Nakajima, S., Miyazaki, J., Uemura, S. Tag-based contextual collaborative filtering. In Int. Journal of Computer Science, IAENG, 34(2), 2007.
8. Niwa, S., Doi, T., Honiden, S. Web page recommender system based on folksonomy mining for ITNG'06 submissions. In Proc.of the 3rd Int. Conf. on Information Tech.: New Generations, Washington, DC, USA, 2006. IEEE, pp. 388-393.
9. Pazzani, M., J., Billsus, D. Content-Based Recommendation Systems. In Brusilovsky P., Kobsa A., Nejdl, W. (eds). LNCS, vol. 4321: The Adaptive Web: Methods and Strategies of Web Personalization. Springer, 2007, pp. 325-341.
10. Šimko, M. Bieliková, M. Automated Educational Course Metadata Generation Based on Semantics Discovery. In LNCS 5794, Proc. of European Conf. on Technology Enhanced Learning, EC TEL 2009, Nice, France. Springer, 2009, pp. 99-105.
11. Šimko,.M., Bieliková, M. Automatic Concept Relationships Discovery for an Adaptive E-course. In Barnes, T. et al. (Eds.). Proc. of Educational Data Mining 2009: 2nd Int. Conf. on Educational Data Mining. Cordoba, Spain, 2009, pp. 171-179.